International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:8, No:10, 2014

Stock Market Prediction by Regression Model with Social Moods

Masahiro Ohmura, Koh Kakusho, Takeshi Okadome

***Abstract*—This paper presents a regression model with autocorrelated errors in which the inputs are social moods obtained by analyzing the adjectives in Twitter posts using a document topic model, where document topics are extracted using LDA. The regression model predicts Dow Jones Industrial Average (DJIA) more precisely than autoregressive moving-average models.

***Keywords*—Regression model, social mood, stock market prediction, Twitter.

I. INTRODUCTION

BOLLEN et al. [1] investigated whether public sentiment, as expressed in large-scale collections of daily Twitterposts, can be used to predict the stock market. They used the Profile of Mood States Bipolar (POMS-bi) Lorr et al.[2] and analysed the text content of tweets to generate a six-dimensional daily time series of public mood ("calm," "alert," "sure," "vital," "kind," and "happy") to provide a more detailed view of changes in the public along a variety of different mood dimensions. They found that the resulting public mood time series were correlated to the Dow Jones Industrial Average (DJIA) to assess their ability to predict changes in the DJIA over time.

Our analysis, however, reveals no correlation between the daily closing prices of the DJIA and the POSM-bi "calm" factor extracted from tweets from the previous day, although Bollen et al. [2] reported a high correlation between them. Instead of using POMS-bi adjectives directly, we construct a method that analyzes tweets with Latent Dirichlet Allocation (LDA) using a set of a day's or half-a-day's worth of tweets as a document [3]. This paper proposes a time-series analysis technique together with daily public mood extracted using the method. The technique predicts the DJIA more precisely than that without the public mood.

II. RELATED WORK

Golder and Macy [4] identified individual-level diurnal and seasonal mood rhythms in cultures across the globe, using data from millions of public Twitter messages. They found that individuals awaken in a good mood that deteriorates as the day progresses (which is consistent with the effects of sleep and circadian rhythm) and that seasonal change in baseline positive affect varies with changes in day length. This may conceivably also be the case for the stock market.

Lee et al. [5] described a novel algorithm called BursT, which uses the sliding-window technique for weighting message streams. BursT facilitates online burst analysis by adopting a long-term expectation of arrival rate as a global baseline. The experimental results showed that the performance of the weighting technique was sufficiently outstanding to reflect the shifts of concept drift, especially when dealing with the issue of diminishing ineffective oral phrases. They concluded that the result of the work can be extended to perform a periodic feature extraction, and they were able to integrate other sophisticated clustering methods to enhance the efficiency for real-time event mining in social networks.

III. PUBLIC MOOD ESTIMATION

This section briefly describes LDA, which is a topic model [6], [7]. Then it summarizes a method that extracts public mood by analyzing tweets using LDA. For detail of the method, see [3].

A. LDA

LDA is a statistical model of document collection that tries to capture this intuition. It is most easily described by its generative process, the imaginary random process by which the model assumes the documents arose. A topic is formally defined as a distribution over a fixed vocabulary. For example, the genetics topic has words about genetics with high probability, and the evolutionary biology topic has words about evolutionary biology with high probability. These topics are assumed to be specified before any data have been generated.

For each document in a collection, LDA generates the words in a two-stage process. That is,

1. Randomly choose a distribution over topics.
2. Randomly choose a topic from the distribution over topics in Step 1.
3. Randomly choose a word from the corresponding distribution over the vocabulary.

More formal description of LDA is as follows.

A) A word is defined to be an item from a vocabulary indexed by $\{1, \ldots,\}$.

B) A document is a sequence of N words denoted by $w = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.

C) A corpus is a collection of M documents denoted by $D = \{w_1, w_2, \ldots, w_M\}$.

Masahiro Ohmura and Koh Kakusho are with the Kwansei Gakuin University, Japan (e-mail: masahiro.ohmura@kwansei.ac.jp, kakusho@kwansei.ac.jp).

Takeshi Okadome is with the Kwansei Gakuin University (corresponding author, e-mail: tokadome@acm.org).

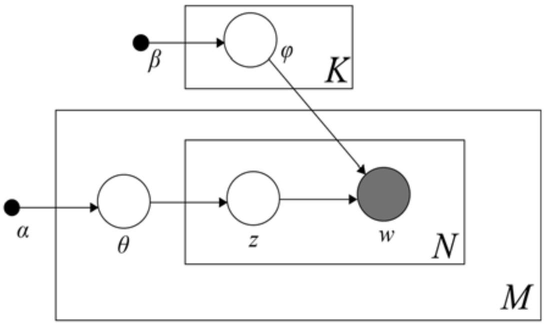

Fig. 1 Graphical model representation of LDA. The boxes are plates representing replicates and the outer plate represents documents, and the inner plate represents the repeated choice of topics and words within a document

Fig. 1 represents a graphical model of LDA. LDA assumes the following productive process for each document $\boldsymbol{w}$ in a corpus D:

For each topic $k = 1, \ldots, K$

I. Draw word distribution,

$$\phi_k \sim Dir(\beta), \beta = (\beta_1, \cdots \beta_V)^T$$

II. Choose $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.

III. For each of the N words w_n:

(a) Choose a topic $z_n \sim$ Multinomial ($\boldsymbol{\theta}$).

(b) Choose a word w_n from $p(w_n | z_n, \boldsymbol{\beta})$, a multinomial probability conditioned on the topic z_n.

For simplicity, the dimensionality K of the Dirichlet distribution (and thus the dimensionality of the topic variable z) is assumed to be known and fixed.

A k-dimensional Dirichlet random variable $\boldsymbol{\theta}$ has the following probability density:

$$p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_k^{\alpha_k - k}, \tag{1}$$

where the parameter $\boldsymbol{\alpha}$ is a k vector with components $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function. Fig. 1 shows the graphical model representation of LDA.

B. Tweet Analysis with LDA

We obtained a collection of public tweets that were recorded from June 28, 2012 to April 30, 2013. After removing stop words and punctuation, we grouped all tweets that were submitted on the same half-day: one from 0:00 AM to 12:00 PM and the other from 12:00 PM to 0:00 AM in EDT.

Along the same line of Bollen et al. [1], we only used tweets that contain explicit statements of their authors' mood states, i.e., those that match the expressions "I feel," "I am feeling," "i'm feeling," "I don't feel," "i'm," "im," "i am," and "makes me." We expand the original lexicons to 800 adjectives by selecting synonyms for them using Word-Net [8]. The enlarged lexicon of 800 terms permits us to capture a wide variety of naturally occurring mood terms in tweets.

We estimate public mood by analyzing tweets with LDA that include at least one of the 800 lexicons. Regarding the collection of half-a-day tweets as a document, we use the topic analysis method with LDA. Because latent topics extracted with LDA reflect word co-occurrences in documents, we expect that "topics" extracted with LDA for tweets including adjectives that express one's feelings correspond to public moods that may reflect the feelings of people posting the tweets. We use Gibbs sampling to compute the posterior distribution.

We evaluate public mood represented with LDA using the method. The method requires specification of the number of public moods, just as the topic analysis with LDA requires a specific number of topics. Because POMS-bi consists of six bipolar scales, each of which has measures of state positive and negative effects, we set the number of public moods to 6 in the evaluation. Table I lists the most frequent words of Moods 1 and 5 that are extracted with LDA when the number of public moods is set to 6.

TABLE I
THE TOP WORDS OF MOODS 1 AND 5 EXTRACTED WITH LDA IN A TEN-MOOD SETTING. FROM [3]

	Mood 1	Mood 5
1	still	know
2	know	get
3	sure	sure
4	love	hard
5	hard	sad
6	school	gonna
7	sorry	people
8	think	school
9	sad	think
10	day	mad

Also Table II shows the average rates of Moods 1 and 5 extracted with LDA in a six-mood setting by using a set of half-a-day's worth of tweets as a document.

TABLE II
THE AVERAGE RATES OF MOODS 1 AND 5 EXTRACTED WITH LDA IN A SIX-MOOD SETTING BY USING A SET OF HALF-A-DAYS' WORTH OF TWEETS AS A DOCUMENT [3]

	Mood 1	Mood 5
Average	0.1513	0.2138

Focusing on the days when great events occurred in the real world, we check whether public mood inferred with LDA reflects real-world trends. That is, we compare public mood extracted from tweets posted on days when a great event occurred with that inferred from tweets posted on days when no great event happened. The results show that one of six public mood extracted by the method reflects positive affect on Thanksgiving Day. Furthermore, another mood captures negative affect on the day of the Boston Marathon bombings.

The evaluation experiment also shows that the measures of the method are insensitive to daily variation of the public mood. This is a disadvantage for predicting social events such as a stock market.

International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:8, No:10, 2014

In order to cope with the disadvantage, we develop a new method, lexicon-reduction method, which is a revised version of the method previously mentioned. That is, by focusing on adjectives that represent feelings in tweets, the method first corresponds each of the adjectives to 1 of 72 representative adjectives, and then it extracts public mood with LDA by using, as a document, a collection of the representative adjectives in daily tweets.

The measures obtained by the lexicon-reduction method are more sensitive to the daily variation of public mood than those obtained by the first method. Table III lists the top words extracted with LDA when the number of public moods is set to 6.

TABLE III
THE TOP WORDS OF EACH MOOD EXTRACTED WITH LDA IN A SIX-MOODS SETTING [3]

	Mood 2	Mood 3	Mood 5
1	mad	downhearted	tired
2	sad	dejected	exhausted
3	nervous	gloomy	weary
4	confused	lively	uneasy
5	perplexed	calm	downhearted

IV. Stock Prediction

We now predict DJIA values as an application of the LR method. Bollen et al. [1] also studied DJIA values the using Granger causality test and self-organizing fuzzy neural network model on the basis of two sets of inputs: (1) the past 3days of DJIA values, and (2) the same, combined with various permutations of their mood time series. However, they did not predict DJIA values, but only the daily up and down changes.

This study predicts the highest prices x_t in the day t using both the highest prices from the beginning to the previous day t -1 and the public mood in the afternoon of the previous day t -1.

For economic indicators such as stock values, we typically look at return y_t (or percentage change) defined by

$$y_t = \frac{x_t - x_{t-1}}{x_t}.$$

A. Regression Model

This study uses a regression model with autocorrelated errors [9], where regressors (fixed inputs) are the mood differences between the afternoons of the previous day and the day before the previous day and autocorrelated errors are expressed by autoregressive (AR) models.

That is, it assumes the following time-series model:

$$y_t = \sum_{i=1}^{k} w_i (d_i^{t-1} - d_i^{t-2}) + \sum_{j=1}^{p} \phi_j y_{t-j} + \varepsilon_t,$$

where d_i^t denotes the ith mood value at t extracted by the LR method, w_i denotes the weight, ϕ_j is the coefficient of AR models, and ε_t is the independent and indentically distributed (i.i.d.) white Gaussian noise. In this study, we use less than or equal to three moods among the six.

From June 30, 2012 to April 19, 2013, the stock market was open 201 days. We select a model using data from the first 60 days, estimate the model parameters using the data of the other 141 days, and predict one-day-ahead stock values for the last 20 days of the 141. That is, using the data of the last 10 days in the first 60 days, we determine the model parameters, the AR model order, the number of regressors, and the public moods, in such a way that the model minimizes the root-mean-square (RMS) error of the one-day-ahead forecast.

We then estimate the weights for the public moods and the coefficients of the AR model using the data from the beginning to each of the last 20 days in the other 141. Finally, we predict one-day-ahead stock values for the last 20 days by the estimated model. In the selected model, the order of AR is second, the number of regressors is 3, and the public moods are Moods 2, 3, and 5. The RMS error of return is 8.58×10^{-4}.

B. Evaluation

We compare the predictions by the regression model with autocorrelated errors with those by autoregressive integrated moving-average (ARIMA) models without fixed regressors of public moods. We determine the autoregressive order p, the integrated order d, and the moving average of order q of the ARIMA(p, d, q) model, in the ranges of $2 \leq p \leq 8$, $0 \leq d \leq 2$, and $0 \leq q \leq 2$, in such a way that the model minimizes the RMS error of one-day-ahead forecast for the last 20 days.

The optimal model is ARIMA (2, 0, 2), or ARMA (2, 2), and the RMS error of return predicted by the ARMA (2, 2) model is 8.94×10^{-4}, which is greater than that by the regression model with autocorrelated errors.

TABLE IV
SUMMARY OF DJIA RMS ERRORS OF RETURN PREDICTED BY THE REGRESSION MODEL WITH AUTOCORRELATED ERRORS AND THAT BY ARMA (2, 2) WITHOUT FIXED REGRESSORS OF PUBLIC MOODS

AR with regressors	ARMA(2, 2)
8.58×10^{-4}	8.94×10^{-4}

Table IV shows the RMS errors of return predicted by the AR model with autocorrelated erros and that by ARMA (2, 2) model without fixed regressors of public moods. The average difference between the RMS error of return predicted by the regression model with autocorrelated errors and that by the ARMA model is 3.6×10^{-6}, which corresponds to approximately 0.5 U.S. dollars (USD) (assuming DJIA is 15,000 USD).

Without discriminating the data for model selection and those for parameter estimation, we also select models and estimate their parameters using the data of the first 181 days in the 201 days and evaluate the forecasts of the models using the data of the last 20 days in the 201. The results are as follows:

- The optimal regression model with autocorrelated errors is selected with Moods 2, 5, and 6 regressors and the fifth order of AR and the RMS error of return predicted by the model is 8.41×10^{-4}.
- The ARMA (5, 2) model is selected among ARIMA (p, d, q) models and the return error predicted by the model is 8.75×10^{-4}.

– The average difference between them is 3.4×10^{-6}.

These results are compatible with the previous ones. Incidentally, the p-values of the Ljung-Box statistic are greater than 0.2, which show that the residuals indicate no patterns and the models we obtained are valid.

V. Discussion

By using, as a document, data from half-a-days' worth of tweets, the method described in Section 3 extracts the social sentiments using LDA from collected tweets that have at least one of 800 sentimental or emotional adjectives. It captures positive public mood and negative public mood, although the other extracted moods do not have intuitive meaning.

The result that many public moods extracted with LDA have no intuitive meaning corresponds to the fact that the topic analysis with LDA extracts many "topics" that have no intuitive meanings, in particular, for documents composed of many articles such as newspapers and magazines that are a mixture of many articles, each of which a different main topic. Thus, public mood changes result in small changes in word co-occurrence. As a result, the method produces slight variations in public moods extracted from half-a-day tweets.

To improve the lowered sensitivity to changes in time, we reduce the 800 adjectives to the corresponding representative 72 adjectives and analyze tweets using LDA for the 72 adjectives with their word frequencies in the half-a-day tweets. The LR method permits us to obtain social sentiments that show improved sensitivity to changes in time. Because the LR method uses only 72 POSM-bi lexicons that are obtained by reducing the 800 adjectives in tweets, small numbers of word co-occurrences accumulate in the reduction and, as a result, reflect on public mood differences between two sets of half-a-day tweets.

Using social sentiments obtained by analyzing the contracted adjectives, we predict DJIA with a regression model with autocorrelated errors, where

1. Fixed inputs are the mood differences between the afternoons of the previous day and the day before the previous day and
2. Autocorrelated errors are expressed by AR models.

The analysis of return predicted by the time-series models shows that the regression model with public moods predicts return better than the ARMA model without them.

VI. Summary

This paper proposed a regression model with autocorrelated errors with the social mood differences as fixed inputs. The social moods are extracted by two methods that analyze the social sentiments from collected tweets that have sentimental or emotional adjectives. The methods infer the public moods with Latent Dirichlet Allocation (LDA) by using half-a-days' worth of tweets as a document. They capture some public moods that match our daily sentiments, although some do not coincide with them.

One of the methods extracts social sentiments that indicate lowered sensitivity to changes in time. By reducing the 800 adjectives to the representative 72 adjectives, the other method enables us to obtain social sentiments that show improved sensitivity to changes in time. The regression model with autocorrelated errors with the mood differences as fixed inputs predicts DJIA value more precisely than ARMA models.

References

[1] Bollen, J., H. Mao, and X.-J. Zeng (2011). Twitter mood predicts the stock market, *Journal of Computational Science*, 2, 1, pp.1-8.
[2] Lorr, M., D. M. McNair, and J. Heuchert (2011). *Poms bi-polar manual supplement*. Multi-Health System Inc.
[3] Ohomura, M., K. Kakusho, T. Okadome (2014). Social mood extraction from Twitter posts with document topic model, *Proceedings of the Fifth International Conference on Information Science and Applications* (*ICISA*2014), 357-360, Soul, May 2014.
[4] Golder, S. A. and M. W. Macy (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333, 6051, pp. 1878-1881.
[5] Lee, C.-H., C.-H. Wu, and T.-F. Chien (2011). BursT: A dynamic term weighting scheme for mining microblogging messages, *Proceedings of the 8th international Conference on Advances in Neural Networks - Volume Part III*, pp. 548-557.
[6] Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, 3, pp. 993-1022.
[7] Blei, D. M. (2012). Probabilistic topic model, *Communications of the Association for Computing Machinery*, 55, 4, pp. 77-84.
[8] Fellbaum, C. and G. Miller (1988). *WordNet: An Electronic Lexical Database* (*Language, Speech, and Communication*). MIT Press.
[9] Shumway, R. H. and D. S. Stoffer (2011). *Time Series Analysis and Its Applications* (Third Edition). Springer, New York.